

LOW-RANK PASSTHROUGH NEURAL NETWORKS

Antonio Valerio Miceli Barone *

School of Informatics
The University of Edinburgh
amiceli@inf.ed.ac.uk

ABSTRACT

Deep learning consists in training neural networks to perform computations that sequentially unfold in many steps over a time dimension or an intrinsic depth dimension. Effective learning in this setting is usually accomplished by specialized network architectures that are designed to mitigate the vanishing gradient problem of naive deep networks. Many of these architectures, such as LSTMs, GRUs, Highway Networks and Deep Residual Network, are based on a single structural principle: the state passthrough.

We observe that these architectures, hereby characterized as Passthrough Networks, in addition to the mitigation of the vanishing gradient problem, enable the decoupling of the network state size from the number of parameters of the network, a possibility that is exploited in some recent works but not thoroughly explored.

In this work we propose simple, yet effective, low-rank and low-rank plus diagonal matrix parametrizations for Passthrough Networks which exploit this decoupling property, reducing the data complexity and memory requirements of the network while preserving its memory capacity. We present competitive experimental results on synthetic tasks and a near state of the art result on sequential randomly-permuted MNIST classification, a hard task on natural data.

1 OVERVIEW

Deep neural networks can perform non-trivial computation by the repeated the application of parametric non-linear transformation layers to vectorial (or, more generally, tensorial) data. This staging of many computation steps can be done over a time dimension for tasks involving sequential inputs or outputs of varying length, yielding a *recurrent neural network*, or over an intrinsic circuit depth dimension, yielding a *deep feed-forward neural network*, or both. Training these deep models is complicated by the *exploding* and *vanishing* gradient problems (Hochreiter, 1991; Bengio et al., 1994).

Starting from the original LSTM of Hochreiter & Schmidhuber (1997), various network architectures have been proposed to ameliorate the vanishing gradient problem in the recurrent neural network setting, such as the modern LSTM (Graves & Schmidhuber, 2005), the GRU (Cho et al., 2014b) and other variants (Greff et al., 2015; Józefowicz et al., 2015). These architectures led to a number of breakthroughs in different tasks such as speech recognition (Graves et al., 2013), machine translation (Cho et al., 2014a; Bahdanau et al., 2014), natural language parsing (Vinyals et al., 2014), question answering (Iyyer et al., 2014) and many others. More recently, similar methods have been applied in the feed-forward neural network setting yielding state of the art results with architectures such as Highway Networks (Srivastava et al., 2015), Deep Residual Networks (He et al., 2015) and Grid LSTM¹ (Kalchbrenner et al., 2015). All these architectures are based on a single structural principle which, in this work, we will refer to as the *state passthrough*. We will thus refer to these architectures as *Passthrough Networks*.

*Work partially done while affiliated with University of Pisa.

¹which also generalize to networks which are deep in both an intrinsic dimension and a time dimension, or even in multiple additional dimensions.

Another difficulty in training neural networks is the trade-off between the network representation power and its number of trainable parameters, which affects its data complexity during training in addition to its implementation memory requirements. More specifically, the number of parameters influences the representation power in two ways: on one hand, it can be thought as the number of tunable "knobs" or "switches" that need to be set to represent a given computable function. On the other hand, however, the number of parameters constrains, in most neural architectures, the size of the partial results that are propagated inside the network: its internal memory capacity.

In typical "fully connected" neural architectures, a layer acting on a n -dimensional state vector has $O(n^2)$ parameters stored in one or more matrices. Since a sufficiently complex function requires a large number of bits to be represented regardless of architectural details, we can't hope to find low-dimensional representation for really hard learning tasks, but there can be many functions of practical interest that are simple enough to be represented by a relatively small number of bits while still requiring some sizable amount of memory to be computed. Therefore, representing these functions on a fully connected neural network can be wasteful in terms of number of parameters. For some tasks, this quadratic dependency between state size and parameter number can cause a model going from underfitting the training set to overfitting it just by the addition of a single state component. For this reason, a number of neural low-dimensional layer parametrization have been proposed, such as convolutional layers (LeCun et al., 2004; Krizhevsky et al., 2012) which impose a sparse, local, periodic structure on the parameter matrices, or multiplicative matrix decompositions, notably the Unitary Evolution RNNs (Arjovsky et al., 2015) (which also addresses the vanishing gradient problem) and others (Le et al., 2013; Moczulski et al., 2015).

In this work we observe that the state passthrough allows for a systematic decoupling of the network state size from the number of parameters: since by default the state vector passes mostly unaltered through the layers, each layer can be made simple enough to be described only by a small number of parameters without affecting the overall memory capacity of the network. This effectively spreads the computation over the depth or time dimension of the network, but without making the network "thin" (as proposed, for instance, by Srivastava et al. (2015)).

To the best of our knowledge, this systematic decoupling has not been described in a systematic way, although it has been exploited by some convolutional passthrough architectures for image recognition (Srivastava et al., 2015; He et al., 2015) or algorithmic tasks (Kaiser & Sutskever, 2015), or architectures with addressable read-write memory (Graves et al., 2014; Gregor et al., 2015; Nee-lakantan et al., 2015; Kurach et al., 2015; Danihelka et al., 2016).

In this work we introduce an unified view of passthrough architectures, describe their state size-parameter size decoupling property, propose simple but effective low-dimensional parametrizations that exploit this decoupling based on low-rank or low-rank plus diagonal matrix decompositions. Our approach extends the LSTM architecture with a single projection layer by Sak et al. (2014) which has been applied to speech recognition, natural language modeling (Józefowicz et al., 2016), video analysis (Sun et al., 2015) et cetera. We provide experimental evaluation of our approach on GRU and Highway Network architectures on various machine learning tasks, including a near state of the art result for the hard task of sequential randomly-permuted MNIST image recognition (Le et al., 2015).

2 MODEL

In this section we will introduce a notation to describe various neural network architectures, then we will formally describe passthrough architectures and finally will introduce our low-dimensional parametrizations for these architectures.

A neural network can be described as a dynamical system that transforms an input u into an output y over multiple time steps T . At each step t the network has a n -dimensional state vector $x(t) \in \mathcal{R}^n$ defined as

$$x(t) = \begin{cases} in(u, \theta) & \text{if } t = 0 \\ f(x(t-1), t, u, \theta) & \text{if } t \geq 1 \end{cases} \quad (1)$$

where in is a *state initialization function*, f is a *state transition function* and $\theta \in \mathcal{R}^k$ is vector of trainable parameters. The output

$$y = out(x(0 : T), \theta) \quad (2)$$

is generated by an *output function* out , where $x(0 : T)$ denotes the whole sequence of states visited during the execution.

In a feed-forward neural network with constant hidden layer width n , the input $u \in \mathcal{R}^m$ and the output $y \in \mathcal{R}^l$ are vectors of fixed dimension m and l respectively, T is a model hyperparameter and the functions above can be simplified as

$$\begin{aligned} in(u, \theta) &= in(u, \theta_{in}) \\ f(x(t-1), t, u, \theta) &= f(x(t-1), \theta_t) \\ out(x(0 : T), \theta) &= out(x(T), \theta_{out}) \end{aligned} \quad (3)$$

highlighting the dependence of the different layers on different subsets of parameters.

In a recurrent neural network the input u is typically a list of T m -dimensional vectors $u(t) \in \mathcal{R}^m$ for $t \in 1, \dots, T$ where T is variable, the output y is either a single l -dimensional vector or a list of T such vectors. The model functions can be written as

$$\begin{aligned} in(u, \theta) &= \theta_{in} \\ f(x(t-1), t, u, \theta) &= f(x(t-1), u(t), \theta_f) \\ out(x(0 : T), \theta) &= [out(x(1), \theta_{out}), \dots, out(x(T), \theta_{out})] \end{aligned} \quad (4)$$

where for a fixed-dimensional output we assume that only $y(T)$ is meaningful.

Other neural architectures, such as "seq2seq" transducers without attention (Cho et al., 2014a), can be also described with this framework.

2.1 PASSTHROUGH NETWORKS

Passthrough networks can be defined as networks where the state transition function f has a special form such that, at each step t the state vector $x(t)$ (or a sub-vector $\hat{x}(t)$) is propagated to the next step modified only by some (nearly) linear, element-wise transformations.

Let the state vector $x(t) \equiv (\hat{x}(t), \tilde{x}(t))$ be the concatenation of $\hat{x}(t) \in \mathcal{R}^{\hat{n}}$ and $\tilde{x}(t) \in \mathcal{R}^{\tilde{n}}$ with $\hat{n} + \tilde{n} = n$ (where \tilde{n} can be equal to zero). We define a network to have a *state passthrough* on \hat{x} if \hat{x} evolves as

$$\hat{x}(t) = f_\pi(x(t-1), t, u, \theta) \odot f_\tau(x(t-1), t, u, \theta) + \hat{x}(t-1) \odot f_\gamma(x(t-1), t, u, \theta) \quad (5)$$

where f_π is the *next state proposal function*, f_τ is the *transform function*, f_γ is the *carry function* and \odot denotes element-wise vector multiplication.

The rest of the state vector $\tilde{x}(t)$, if present, evolves according to some other function \tilde{f} . In practice $\tilde{x}(t)$ is only used in LSTM variants, while in other passthrough architectures $\tilde{x}(t) = x(t)$.

We denote the state passthrough as *additive* if $f_\tau(x(t-1), t, u, \theta) = f_\gamma(x(t-1), t, u, \theta) = 1^{\otimes \hat{n}}$. This choice is used in the original LSTM of Hochreiter & Schmidhuber (1997) and in the Deep Residual Network² of He et al. (2015).

We denote the state passthrough as *convex* if $f_\tau(x(t-1), t, u, \theta) = 1^{\otimes \hat{n}} - f_\gamma(x(t-1), t, u, \theta)$. This choice is used in GRUs (Cho et al., 2014b) and Highway Networks (Srivastava et al., 2015). Modern LSTM variants (Greff et al., 2015) typically use a transform function ("forget gate") f_τ and carry function ("input gate") f_γ independent of each other.

As concrete example, we can describe a fully connected Highway Network as

$$\begin{aligned} f_\pi(x(t-1), t, u, \theta) &= g(\theta_t^{(W_\pi)} \cdot x(t-1) + \theta_t^{(b_\pi)}) \\ f_\tau(x(t-1), t, u, \theta) &= \sigma(\theta_t^{(W_\tau)} \cdot x(t-1) + \theta_t^{(b_\tau)}) \\ f_\gamma(x(t-1), t, u, \theta) &= 1^{\otimes n} - f_\tau(x(t-1), t, u, \theta) \end{aligned} \quad (6)$$

where g is an element-wise activation function, usually the ReLU (Glorot et al., 2011) or the hyperbolic tangent, σ is the element-wise logistic sigmoid, and $\forall t \in 1, \dots, T$, the parameters $\theta_t^{(W_\pi)}$

²the Deep Residual Network does not exactly fit this definition of passthrough network due to the ReLU non-linearities applied between the layers, but it is similar enough that it can be considered to be based on the same principle

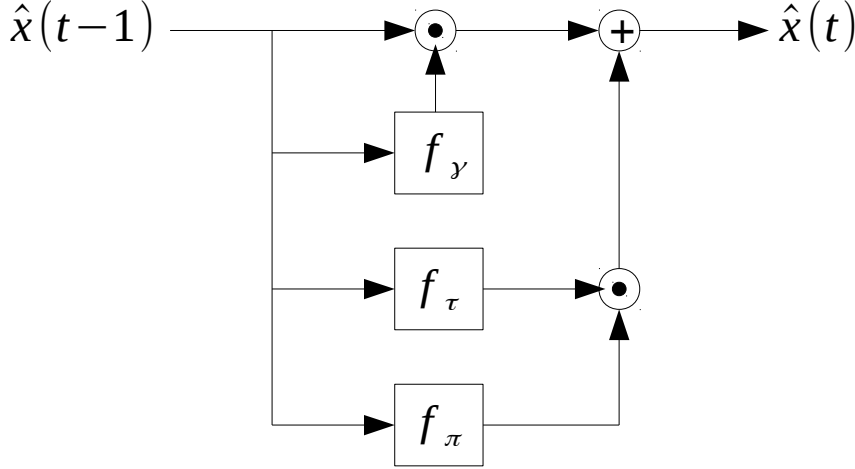


Figure 1: Generic state passthrough hidden layer. Optional non-passthrough state $\tilde{x}(t)$ and per-timestep input $u(t)$ are not shown.

and $\theta_t^{(W_\tau)}$ are matrices in $\mathcal{R}^{n \times n}$ and $\theta_t^{(b_\pi)}$ and $\theta_t^{(b_\tau)}$ are vectors in \mathcal{R}^n . Dependence on the input u occurs only through the initialization function, which is model-specific and is omitted here, as is the output function.

2.2 LOW-RANK PASSTHROUGH NETWORKS

In fully connected architectures there are $n \times n$ matrices that act on the state vector, such as the $\theta_t^{(W_\pi)}$ and $\theta_t^{(W_\tau)}$ matrices of the Highway Network of eq. 6. Each of these matrices has n^2 entries, thus for large n , the entries of these matrices can make up the majority of independently trainable parameters of the model.

As discussed in the previous section, this parametrization can be wasteful. Specifically, this parametrization implies that, at each step, all the information in each state component can affect all the information in any state component at the next step. That is, the computation performed at each step is essentially fully global. Classical physical systems, however, consist of spatially separated parts with primarily local interactions, long-distance interactions are possible but they tend to be limited by propagation delays, bandwidth and noise. Therefore it may be beneficial to bias our model class towards models that tend to adhere to these physical constraints by using a parametrization which reduces the number of parameters required to represent them.

We can accomplish this low-dimensional parametrization by imposing some constraints on the $n \times n$ matrices that parametrize the state transitions. One way of doing this is to impose a convolutional structure on these matrices, which corresponds to strict locality and periodicity constraints as in a cellular automaton. These constraints may work well in certain domains such as vision, but may be overly restrictive in other domains.

We propose instead to impose a low-rank constraint on these matrices. This is easily accomplished by rewriting each of these matrices as the product of two matrices where the inner dimension d is a model hyperparameter. For instance, in the case of the Highway Network of eq. 6 we can redefine $\forall t \in 1, \dots, T$

$$\begin{aligned}\theta_t^{(W_\pi)} &= \theta_t^{(L_\pi)} \cdot \theta_t^{(R_\pi)} \\ \theta_t^{(W_\tau)} &= \theta_t^{(L_\tau)} \cdot \theta_t^{(R_\tau)}\end{aligned}\tag{7}$$

where $\theta_t^{(L_\pi)}, \theta_t^{(L_\tau)} \in \mathcal{R}^{n \times d}$ and $\theta_t^{(R_\pi)}, \theta_t^{(R_\tau)} \in \mathcal{R}^{d \times n}$. When $d < n/2$ this results in a reduction of the number of independent parameters of the model.

This low-rank constraint can be thought of as a bandwidth constraint on the computation performed at each step: the R matrices first project the state into a smaller subspace, extracting the information

needed for that specific step, then the L matrices project it back to the original state space, spreading the selected information to all the state components that need to be updated.

Note that if we were to apply this constraint to a non-passthrough architecture, such as a Multi-Layer Perceptron or a Elman’s Recurrent Neural Network, it would create an information bottleneck within each layer, effectively reducing the memory capacity of the model. But in a passthrough architecture the memory capacity is unaffected since the state passthrough takes care of propagating all the information that does not need to be updated during one step to the next step. Therefore we exploit the decoupling property of the state passthrough. A similar approach has been proposed for the LSTM architecture by Sak et al. (2014), although they force the the R matrices to be the same for all the functions of the state transition, while we allow each parameter matrix to be parametrized independently by a pair of R and L matrices.

Low-rank passthrough architectures are universal in that they retain the same representation classes of their parent architectures. This is trivially true if the inner dimension d is allowed to be $O(n)$ in the worst case, and for some architectures even if d is held constant. For instance, it is easily shown that for any Highway Network with state size n and T hidden layers and for any $\epsilon > 0$, there exist a Low-rank Highway Network with $d = 1$, state size at most $2n$ and at most nT layers that computes the same function within an ϵ margin of error.

2.3 LOW-RANK PLUS DIAGONAL PASSTHROUGH NETWORKS

As we show in the experimental section, on some tasks the low-rank constraint may prove to be excessively restrictive if the goal is to train a model with fewer parameters than one with arbitrary matrices. A simple extension is to add to each low-rank parameter matrix a diagonal parameter matrix, yielding a matrix that is full-rank but still parametrized in a low-dimensional space. For instance, for the Highway Network architecture we modify eq. 7 to

$$\begin{aligned}\theta_t^{(W_\pi)} &= \theta_t^{(L_\pi)} \cdot \theta_t^{(R_\pi)} + \theta_t^{(D_\pi)} \\ \theta_t^{(W_\tau)} &= \theta_t^{(L_\tau)} \cdot \theta_t^{(R_\tau)} + \theta_t^{(D_\tau)}\end{aligned}\tag{8}$$

where $\theta_t^{(D_\pi)}, \theta_t^{(D_\tau)} \in \mathcal{R}^{n \times n}$ are trainable diagonal parameter matrices.

Low-rank plus diagonal decompositions have been used for over a century in factor analysis in statistics (Spearman, 1904), system identification (Kalman, 1982) and other applications. They arise naturally in the estimation of linear relationships between variables from noisy measurements, under certain independence assumptions on the measurement noise. Refer to Saunderson et al. (2012) and Ning et al. (2015) for a review.

At first, it may seem that adding diagonal parameter matrices is redundant in passthrough networks. After all, the state passthrough itself can be considered as a diagonal matrix applied to the state vector, which is then additively combined to the new proposed state computed by the f_π function. However, since the state passthrough completely skips over all non-linear activation functions (except in the Residual Network architecture where it only skips over some of them), these formulations are not equivalent. In particular, the low-rank plus diagonal parametrization may help in recurrent neural networks which receive input at each time step, since they allow each component of the state vector to directly control how much input signal is inserted into it at each step. We demonstrate the effectiveness of this model in the sequence copy tasks described in the experiments section.

3 EXPERIMENTS

In this section we report a preliminary experiment on Low-rank Highway Networks on the MNIST dataset and several experiments on Low-rank GRUs.

3.1 LOW-RANK HIGHWAY NETWORKS

We applied the low-rank and low-rank plus diagonal Highway Network architecture to the classic benchmark task of handwritten digit classification on the MNIST dataset.

We used the low-rank architecture described by equations 6 and 7, with $T = 5$ hidden layers, ReLU activation function, state dimension $n = 1024$ and maximum rank (internal dimension) $d = 256$.

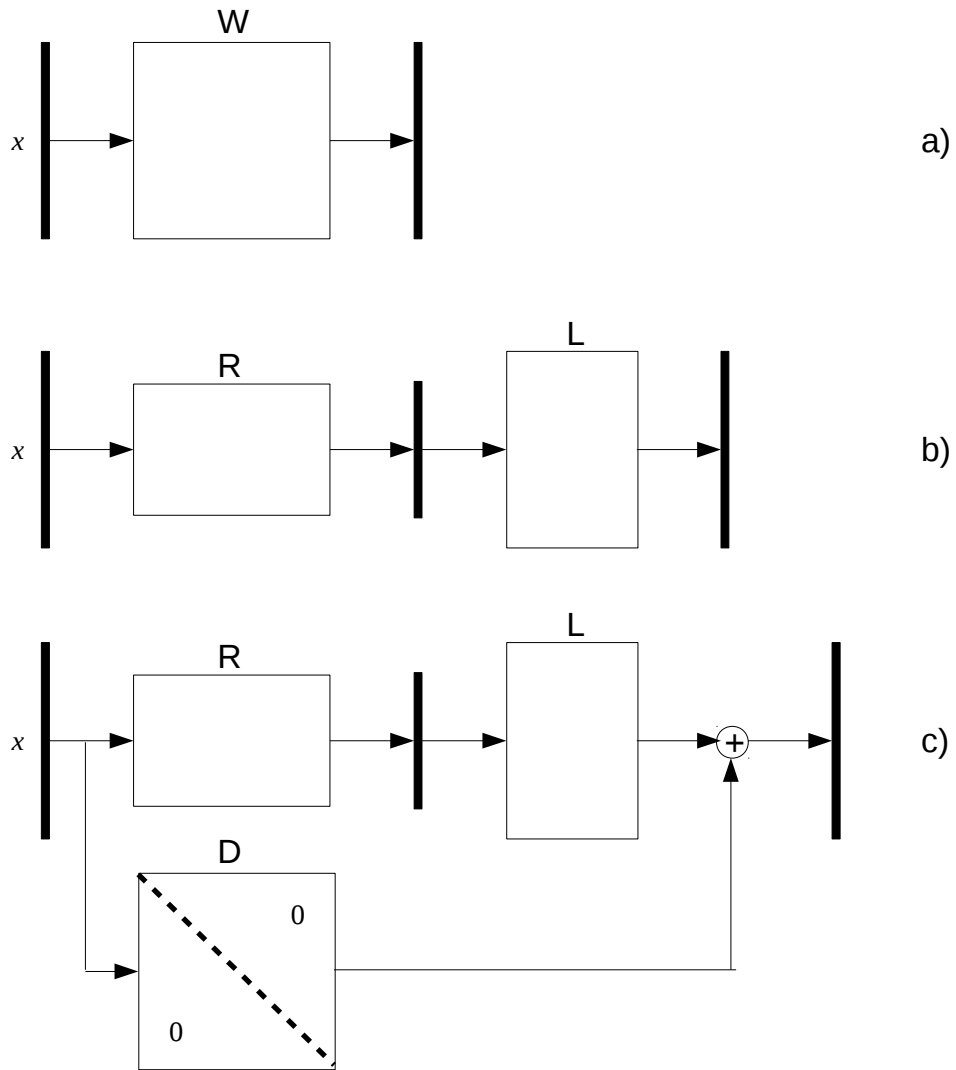


Figure 2: a) Full matrix parametrization. b) Low-rank parametrization. c) Low-rank plus diagonal parametrization.

The input-to-state layer is a dense 784×1024 matrix followed by a (biased) ReLU activation and the state-to-output layer is a dense 1024×10 matrix followed by a (biased) identity activation. We did not use any convolution layer, pooling layer or data augmentation technique.

We used dropout (Srivastava et al., 2014) in order to achieve regularization. We applied standard dropout layers with dropout probability $p = 0.2$ just before the input-to-state layer and $p = 0.5$ just before the state-to-output layer. We also applied dropout inside each hidden layer in the following way: we inserted dropout layers with $p = 0.3$ inside both the proposal function and the transform function, immediately before both the R matrices and the L matrices, totaling to four dropout layers per hidden layer, although the random dropout matrices are shared between proposal and transform functions. Dropout applied this way does not disrupt the state passthrough, thus it does not cause a reduction of memory capacity during training. We further applied L2-regularization with coefficient $\lambda = 1 \times 10^{-3}$ per example on the hidden-to-output parameter matrix.

We also used batch normalization (Ioffe & Szegedy, 2015) after the input-to-state matrix and after each parameter matrix in the hidden layers.

Parameter matrices are randomly initialized using an uniform distribution with scale equal to $\sqrt{6/a}$ where a is the input dimension. Initial bias vectors are all initialized at zero except for those of the transform functions in the hidden layers, which are initialized at -1.0 .

We trained to minimize the sum of the per-class L2-hinge loss plus the L2-regularization cost (Tang, 2013). Optimization was performed using Adam (Kingma & Ba, 2014) with standard hyperparameters, learning rate starting at 3×10^{-3} halving every three epochs without validation improvements. Mini-batch size was equal to 100. Code is available online³.

We ran our experiments on a machine with a 24 core Intel(R) Xeon(R) CPU X5670 2.93GHz, 24 GB of RAM. We did not use a GPU. Training took approximately 4 hours .

We obtained perfect training accuracy and 98.83% test accuracy. While this result does not reach the state of the art for this task (99.13% test accuracy with unsupervised dimensionality reduction reported by Tang (2013)), it is still relatively close.

We also tested the low-rank plus diagonal Highway Network architecture of eq. 8 with the same settings as above, obtaining a test accuracy of 98.64%. The inclusion of diagonal parameter matrices does not seem to help in this particular task.

3.2 LOW-RANK GRUS

We applied the Low-rank and Low-rank plus diagonal GRU architectures to a subset of sequential benchmarks described in the Unitary Evolution Recurrent Neural Networks article by Arjovsky et al. (2015), specifically the memory task, the addition task and the sequential randomly permuted MNIST task. For the memory tasks, we also considered two different variants proposed by Danihelka et al. (2016) and Henaff et al. (2016) which are hard for the uRNN architecture.

We chose to compare against the uRNN architecture because it set state of the art results in terms of both data complexity and accuracy and because it is an architecture with similar design objectives as low-rank passthrough architectures, namely a low-dimensional parametrization and the mitigation of the vanishing gradient problem, but it is based on quite different principles (it does not use a state passthrough as defined in this work, instead it relies on the reversibility and norm-preservation properties of unitary matrices in order preserve state information between time steps, and uses a multiplicative unitary decomposition in order to achieve low-dimensional parametrization).

The GRU architecture (Cho et al., 2014b) is a passthrough recurrent neural network defined as

$$\begin{aligned}
in(u, \theta) &= \theta_{in} \\
f_{\omega}(x(t-1), t, u, \theta) &= \sigma(\theta^{U_{\omega}} \cdot u(t) + \theta^{(W_{\omega})} \cdot x(t-1) + \theta^{(b_{\omega})}) \\
f_{\gamma}(x(t-1), t, u, \theta) &= \sigma(\theta^{U_{\gamma}} \cdot u(t) + \theta^{(W_{\gamma})} \cdot x(t-1) + \theta^{(b_{\gamma})}) \\
f_{\pi}(x(t-1), t, u, \theta) &= 1^{\otimes n} - f_{\gamma}(x(t-1), t, u, \theta) \\
f_{\pi}(x(t-1), t, u, \theta) &= g(\theta^{U_{\pi}} \cdot u(t) + \theta^{(W_{\pi})} \cdot (x(t-1) \odot f_{\omega}(x(t-1), t, u, \theta)) + \theta^{(b_{\pi})})
\end{aligned} \tag{9}$$

³<https://github.com/Avmb/lowrank-highwaynetwork>

Note that with respect of the definition of the Highway Network architecture of eq. 6, the initial state θ_{in} is a model parameter, there is an additional function f_ω (the "reset" gate), parameters don't depend on time t and input $u(t)$ is included in the computation at each step though the θ^U matrices. We have also defined the transform function f_τ in terms of the carry function f_γ rather than vice versa for consistency with the literature, although the two formulations are isomorphic.

We turn this architecture into the Low-rank GRU architecture by redefining each of the θ^W matrices as the product of two matrices with inner dimension d . For the memory tasks, which turned out to be difficult for the low-rank parametrization, we also consider the low-rank plus diagonal parametrization. We also applied the low-rank plus diagonal parametrization for the sequential permuted MNIST task.

In our experiments we optimized using RMSProp (Tieleman & Hinton, 2012) with gradient component clipping at 1. Code is available online⁴. Our code is based on the published uRNN code⁵ (specifically, on the LSTM implementation) by the original authors for the sake of a fair comparison. In order to achieve convergence on the memory task however, we had to slightly modify the optimization procedure, specifically we changed gradient component clipping with gradient norm clipping (with NaN detection and recovery), and we added a small $\epsilon = 1 \times 10^{-8}$ term in the parameter update formula. No modifications of the original optimizer implementation were required for the other tasks.

We ran our experiments on the same machine as the experiments described in the previous section, with the exception of the largest sequential permuted MNIST experiment (low-rank plus diagonal GRU with $n = 256$, $d = 24$ which was run on a machine with a Geforce GTX TITAN X GPU).

We will now present a short description of each task, the experimental details and results.

3.2.1 MEMORY TASK

The input of an instance of this task is a sequence of $T = N + 20$ discrete symbols in a ten symbol alphabet $a_i : i \in 0, \dots, 9$, encoded as one-hot vectors. The first 10 symbols in the sequence are "data" symbols i.i.d. sampled from a_0, \dots, a_7 , followed by $N - 1$ "blank" a_8 symbols, then a distinguished "run" symbol a_9 , followed by 10 more "blank" a_8 symbols. The desired output sequence consists of $N + 10$ "blank" a_8 symbols followed by the 10 "data" symbols as they appeared in the input sequence. Therefore the model has to remember the 10 "data" symbol string over the temporal gap of size N , which is challenging for a recurrent neural network when N is large. In our experiment we set $N = 500$, which is the hardest setting explored in the uRNN work. The training set consists of 100,000 training examples and 10,000 validation/test examples.

The architecture is described by eq. (9), with an additional output layer with a dense $n \times 10$ matrix followed a (biased) softmax. We train to minimize the cross-entropy loss.

We were able to solve this task using a GRU with full recurrent matrices with state size $n = 128$, learning rate 1×10^{-3} , mini-batch size 20, initial bias of the carry functions (the "update" gates) 4.0, however this model has many more parameters, nearly 50,000 in the recurrent layer only, than the uRNN work which has about 6,500, and it converges much more slowly than the uRNN.

We were not able to achieve convergence with a pure low-rank model without exceeding the number of parameters of the fully connected model, but we achieved fast convergence with a low-rank plus diagonal model with $d = 50$, with other hyperparameters set as above. This model has still more parameters (39,168 in the recurrent layer, 41,738 total) than the uRNN model and converges more slowly but still reasonably fast, reaching test cross-entropy $< 1 \times 10^{-3}$ nats and almost perfect classification accuracy in less than 35,000 updates.

We also consider two variants of this task which are difficult for the uRNN model. For both these tasks we used the same settings as above except that the task size parameter is set at $N = 100$ for consistency with the works that introduced these variants.

In the variant of Danihelka et al. (2016), the length of the sequence to be remembered is randomly sampled between 1 and 10 for each sequence. They manage to achieve fast convergence with

⁴<https://github.com/Avmb/lowrank-gru>

⁵https://github.com/amarshah/complex_rnn

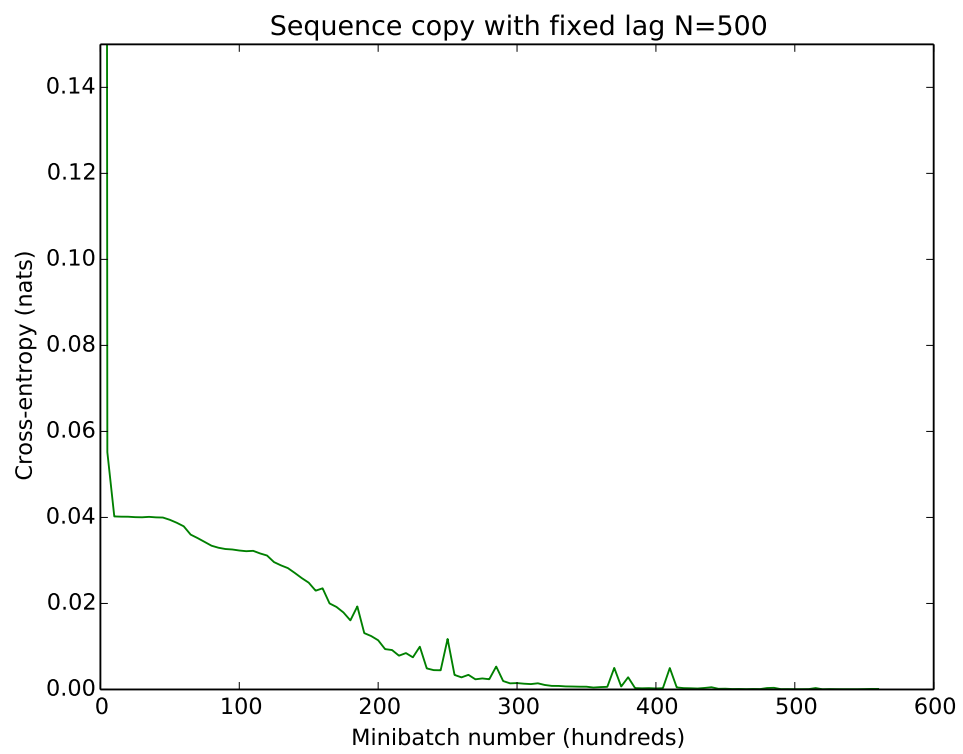


Figure 3: Low-rank plus diagonal GRU on the fixed-length sequence copy task with fixed lag. Cross-entropy on validation set.

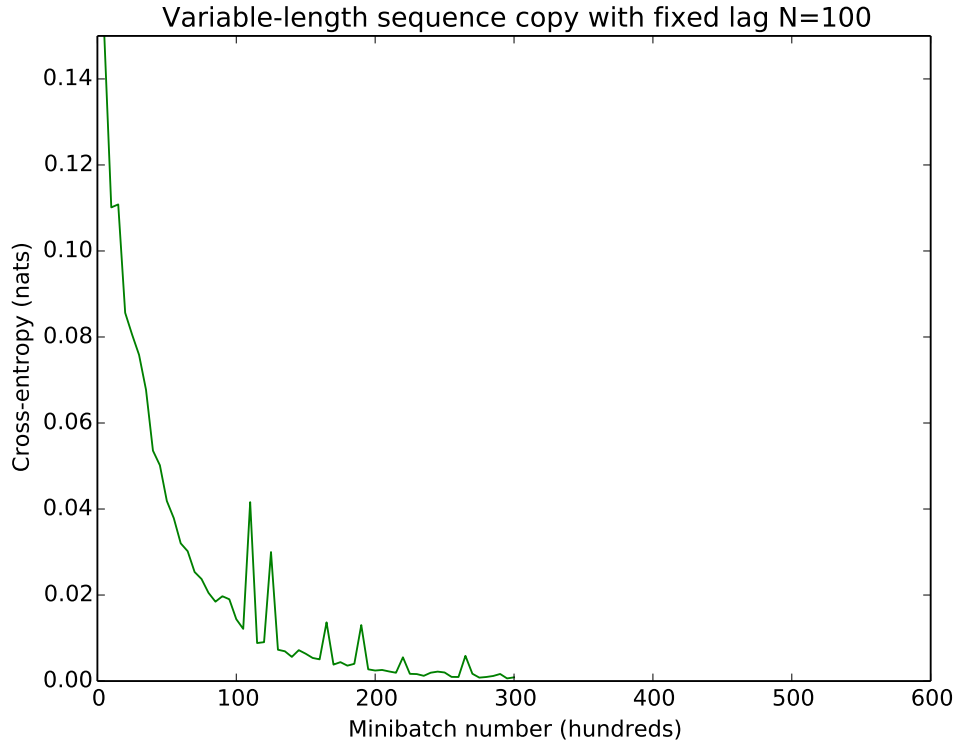


Figure 4: Low-rank plus diagonal GRU on the variable-length sequence copy task with fixed lag. Cross-entropy on validation set.

their Associative LSTM architecture with 65,505 parameters, and slower convergence with standard LSTM models. Our low-rank plus diagonal GRU architecture, which has less parameters than their Associative LSTM, performs comparably or better, reaching test cross-entropy $< 1 \times 10^{-3}$ nats and almost perfect classification accuracy in less than 30,000 updates.

In the variant of Henaff et al. (2016), the length of the sequence to be remembered is fixed at 10 but the model is expected to copy it after a variable number of time steps randomly chosen, for each sequence, between 1 and $N = 100$. The authors achieve slow convergence with a standard LSTM model, while our low-rank plus diagonal GRU architecture achieves fast convergence, reaching test cross-entropy $< 1 \times 10^{-3}$ nats and almost perfect classification accuracy in less than 38,000 updates, and perfect test accuracy in 87,000 updates.

3.2.2 ADDITION TASK

For each instance of this task, the input sequence has length T and consists of two real-valued components, at each step the first component is independently sampled from the interval $[0, 1]$ with uniform probability, the second component is equal to zero everywhere except at two randomly chosen time step, one in each half of the sequence, where it is equal to one. The result is a single real value computed from the final state which we want to be equal to the sum of the two elements of the first component of the sequence at the positions where the second component was set at one. In our experiment we set $T = 750$. The training set consists of 100,000 training examples and 10,000 validation/test examples.

We use a Low-rank GRU with $2 \times n$ input matrix, $n \times 1$ output matrix and (biased) identity output activation. We train to minimize the mean squared error loss. We use the following hyperparameter configuration: State size $n = 128$, maximum rank $d = 24$. This results in approximately 6,140

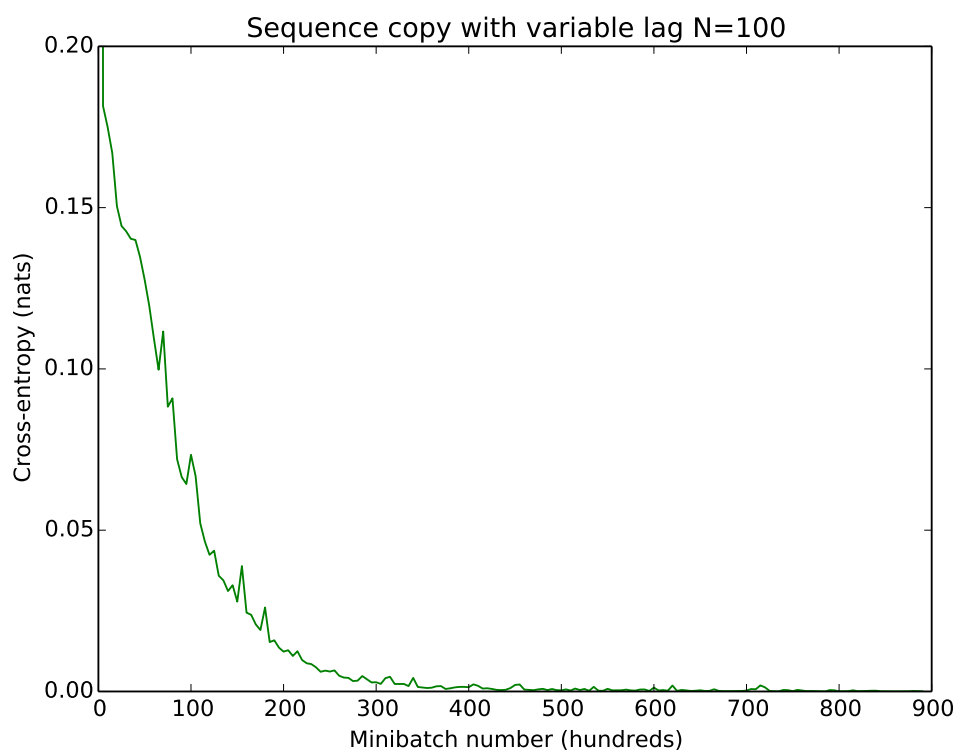


Figure 5: Low-rank plus diagonal GRU on the fixed-length sequence copy task with variable lag. Cross-entropy on validation set. (note that the axes scale is different than fig. 3 and 4.)

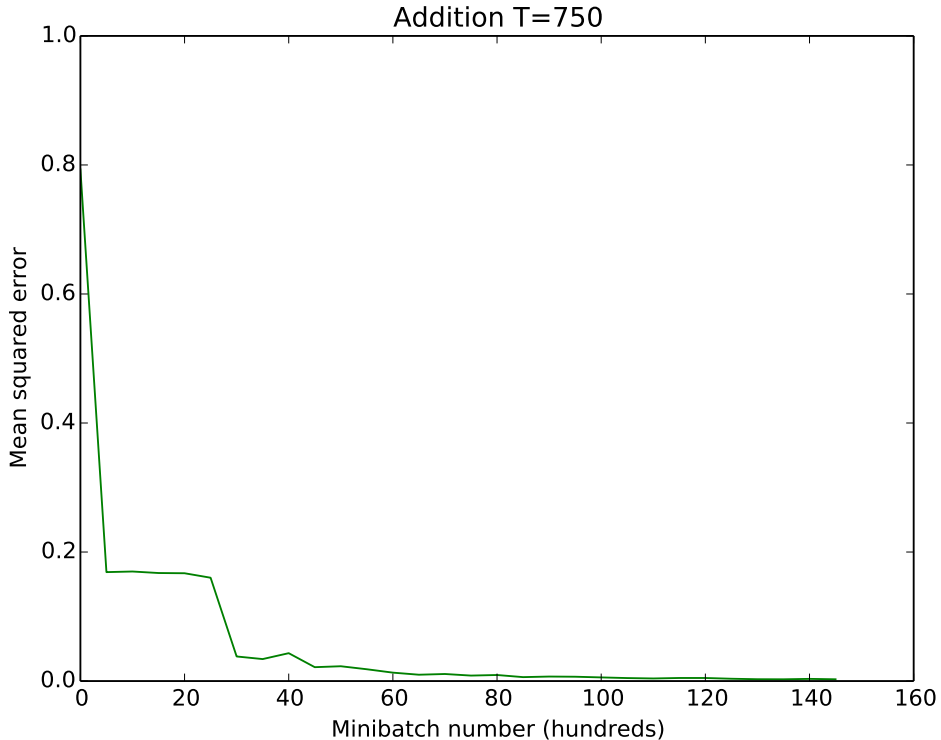


Figure 6: Low-rank GRU on the addition task. Mean squared error on validation set.

parameters in the recurrent hidden layer. Learning rate was set at 1×10^{-3} , mini-batch size 20, initial bias of the carry functions (the "update" gates) was set to 4.

We trained on 14,500 mini-batches, obtaining a mean squared error on the test set of 0.003, which is a better result than the one reported in the uRNN article, in terms of training time and final accuracy.

3.2.3 SEQUENTIAL MNIST TASK

This task consists of handwritten digit classification on the MNIST dataset with the caveat that the input is presented to the model one pixel value at time, over $T = 784$ time steps. To further increase the difficulty of the task, the inputs are reordered according to a random permutation (fixed for all the task instances).

We use a Low-rank GRU with $1 \times n$ input matrix, $n \times 10$ output matrix and (biased) softmax output activation.

Learning rate was set at 5×10^{-4} , mini-batch size 20, initial bias of the carry functions (the "update" gates) was set to 5.

We considered two hyperparameter configurations:

1. State size $n = 128$, maximum rank $d = 24$.
2. State size $n = 512$, maximum rank $d = 4$.

Configuration 1 reaches a validation accuracy of 93.4% in 320,400 iterations. Final test accuracy is 91.8%. The reported uRNN accuracy is 91.4%. Our model however takes 100,500 to reach a validation accuracy comparable to the final accuracy of the uRNN model, which is instead reached in about 20,000 iterations.

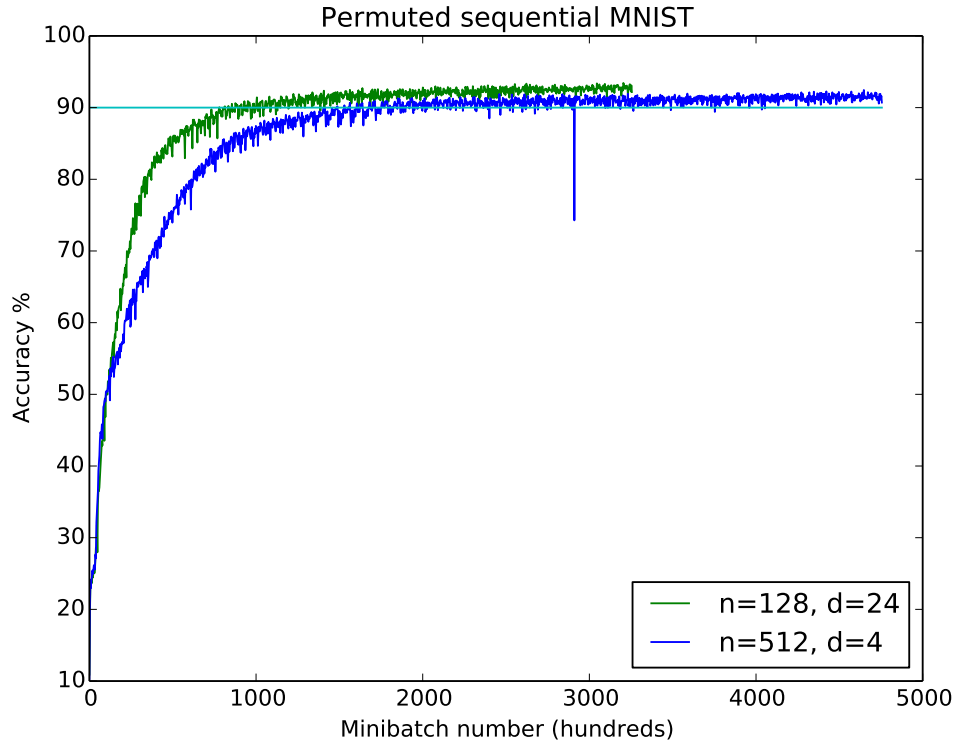


Figure 7: Low-rank GRU on the permuted sequential MNIST task. Accuracy on validation set. Horizontal line indicates 90% accuracy.

Configuration 2 reaches a validation accuracy of 92.5% in 464,700 iterations, with test accuracy of 91.3%. Note that even with the rather extreme bottleneck of $d = 4$, this model performs well.

For this task, we also consider three low-rank plus diagonal parametrizations. We report the best validation accuracy and test accuracy results, in addition to the results for a full-rank baseline GRU:

1. State size $n = 64$, maximum rank $d = 24$. Validation accuracy: 93.1%, test accuracy: 91.9%.
2. State size $n = 128$, maximum rank $d = 24$. Validation accuracy: 94.1%, test accuracy: 93.5%.
3. State size $n = 128$, full-rank. Validation accuracy: 93.0%, test accuracy: 92.8%.
4. State size $n = 256$, maximum rank $d = 24$. Validation accuracy: 95.1%, test accuracy: 94.7%.

Note that the low-rank plus diagonal GRU is more accurate than the full rank GRU with the same state size, while the low-rank GRU is slightly less accurate, indicating the utility of the diagonal component of the parametrization for this task.

These results surpass the uRNN and are on par with more complex architectures with time-skip connections (Zhang et al., 2016) (reported test set accuracy 94.0%). To our knowledge, at the time of this writing, the best result on this task is the LSTM with recurrent batch normalization by Cooijmans et al. (2016) (reported test set accuracy 95.2%). The architectural innovations of these works are orthogonal to our own and in principle they can be combined to it.

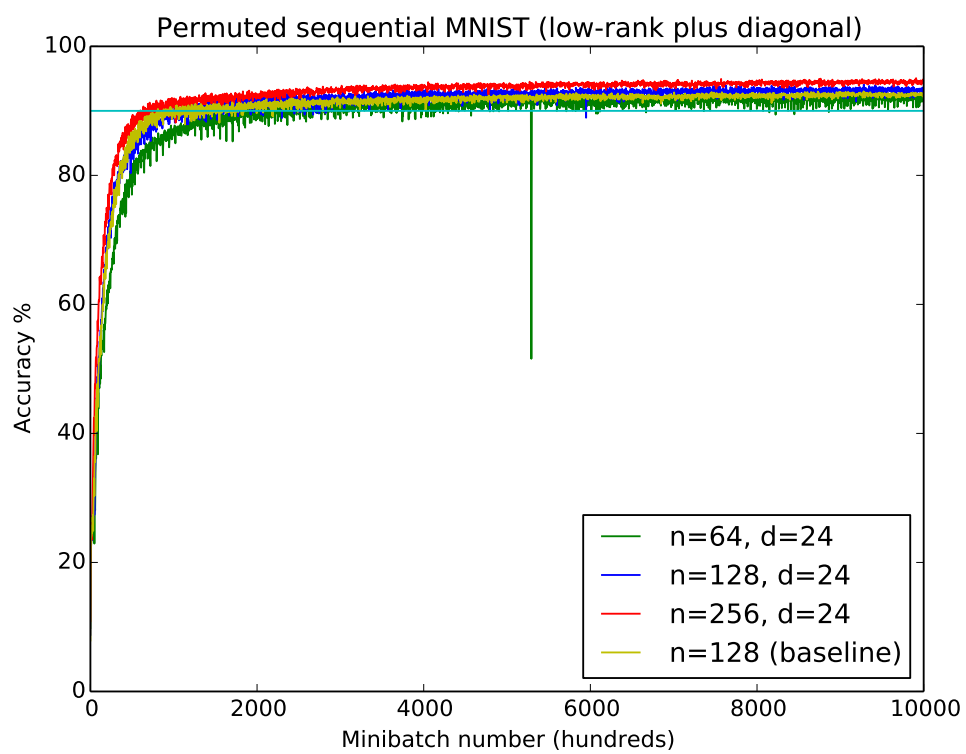


Figure 8: Low-rank plus diagonal GRU and baseline GRU on the permuted sequential MNIST task. Accuracy on validation set.

4 CONCLUSIONS AND FUTURE WORK

We presented a framework that unifies the description various types of recurrent and feed-forward neural networks as passthrough neural networks.

We proposed low-dimensional parametrizations for passthrough neural networks based on low-rank or low-rank plus diagonal decompositions of the $n \times n$ matrices that occur in the hidden layers.

We experimentally compared our models with state of the art models, obtaining competitive results including a state of the art for the randomly-permuted sequential MNIST task.

Our parametrizations are alternative to convolutional parametrizations explored by Srivastava et al. (2015); He et al. (2015); Kaiser & Sutskever (2015). We note that the two approaches can be combined in at least two ways:

- A low-rank (plus diagonal) decomposition (with a suitable axis reshaping) can be applied to convolutional filter banks when the number of channels is large.
- The "local" state acted on by the convolutional passthrough filters can be paired with a "global" state acted on by low-rank (plus diagonal) passthrough matrices. The global state is replicated on additional channels to update the local state and the local state is pooled to update the global state. This arrangement may be useful in particular in the Neural GPU (Kaiser & Sutskever, 2015) in order to augment the cellular automaton with "global variables", which would otherwise need to be replicated on the cell states and threaded over the computation.

Low-rank and low-rank plus diagonal parametrizations are linear, alternative parametrizations could include non-linear activation functions, effectively replacing each hidden parameter matrix with a MLP, similar to the network-in-network approach of Lin et al. (2013).

We leave the exploration of these extensions to future work.

ACKNOWLEDGMENTS

We thank Giuseppe Attardi and the Department of Computer Science of University of Pisa for letting us use their machines to run the experiments presented in this paper.

REFERENCES

- Arjovsky, Martin, Shah, Amar, and Bengio, Yoshua. Unitary evolution recurrent neural networks. *CoRR*, abs/1511.06464, 2015. URL <http://arxiv.org/abs/1511.06464>.
- Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL <http://arxiv.org/abs/1409.0473>.
- Bengio, Yoshua, Simard, Patrice, and Frasconi, Paolo. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994.
- Cho, Kyunghyun, van Merriënboer, Bart, Bahdanau, Dzmitry, and Bengio, Yoshua. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014a.
- Cho, Kyunghyun, van Merriënboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014b.
- Cooijmans, T., Ballas, N., Laurent, C., Gülçehre, Ç., and Courville, A. Recurrent Batch Normalization. *ArXiv e-prints*, March 2016.
- Danihelka, I., Wayne, G., Uria, B., Kalchbrenner, N., and Graves, A. Associative Long Short-Term Memory. *ArXiv e-prints*, February 2016.

-
- Glorot, Xavier, Bordes, Antoine, and Bengio, Yoshua. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.
- Graves, Alex and Schmidhuber, Jürgen. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.
- Graves, Alex, Mohamed, Abdel-rahman, and Hinton, Geoffrey E. Speech recognition with deep recurrent neural networks. *CoRR*, abs/1303.5778, 2013. URL <http://arxiv.org/abs/1303.5778>.
- Graves, Alex, Wayne, Greg, and Danihelka, Ivo. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.
- Greff, Klaus, Srivastava, Rupesh Kumar, Koutník, Jan, Steunebrink, Bas R, and Schmidhuber, Jürgen. Lstm: A search space odyssey. *arXiv preprint arXiv:1503.04069*, 2015.
- Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- Henaff, M., Szlam, A., and LeCun, Y. Orthogonal RNNs and Long-Memory Tasks. *ArXiv e-prints*, February 2016.
- Hochreiter, Sepp. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 1991.
- Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Iyyer, Mohit, Boyd-Graber, Jordan, Claudino, Leonardo, Socher, Richard, and Daumé III, Hal. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 633–644, 2014.
- Józefowicz, Rafal, Zaremba, Wojciech, and Sutskever, Ilya. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 2342–2350, 2015. URL <http://jmlr.org/proceedings/papers/v37/jozefowicz15.html>.
- Józefowicz, Rafal, Vinyals, Oriol, Schuster, Mike, Shazeer, Noam, and Wu, Yonghui. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.
- Kaiser, Lukasz and Sutskever, Ilya. Neural gpu learn algorithms. *CoRR*, abs/1511.08228, 2015. URL <http://arxiv.org/abs/1511.08228>.
- Kalchbrenner, Nal, Danihelka, Ivo, and Graves, Alex. Grid long short-term memory. *arXiv preprint arXiv:1507.01526*, 2015.
- Kalman, R.E. et al. *System Identification from Noisy Data*. Defense Technical Information Center, 1982. URL <https://books.google.it/books?id=T-dCNwAACAAJ>.
- Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- Kurach, Karol, Andrychowicz, Marcin, and Sutskever, Ilya. Neural random-access machines. *CoRR*, abs/1511.06392, 2015. URL <http://arxiv.org/abs/1511.06392>.

-
- Le, Quoc, Sarlós, Tamás, and Smola, Alex. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, 2013.
- Le, Quoc V, Jaitly, Navdeep, and Hinton, Geoffrey E. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015.
- LeCun, Yann, Huang, Fu Jie, and Bottou, Leon. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pp. II–97. IEEE, 2004.
- Lin, Min, Chen, Qiang, and Yan, Shuicheng. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- Moczulski, Marcin, Denil, Misha, Appleyard, Jeremy, and de Freitas, Nando. ACDC: A structured efficient linear layer. *CoRR*, abs/1511.05946, 2015. URL <http://arxiv.org/abs/1511.05946>.
- Neelakantan, Arvind, Le, Quoc V., and Sutskever, Ilya. Neural programmer: Inducing latent programs with gradient descent. *CoRR*, abs/1511.04834, 2015. URL <http://arxiv.org/abs/1511.04834>.
- Ning, Lipeng, Georgiou, Tryphon T, Tannenbaum, Allen, and Boyd, Stephen P. Linear models based on noisy data and the frisch scheme. *SIAM Review*, 57(2):167–197, 2015.
- Sak, Hasim, Senior, Andrew W, and Beaufays, Françoise. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*, pp. 338–342, 2014.
- Saunderson, James, Chandrasekaran, Venkat, Parrilo, Pablo A, and Willsky, Alan S. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.
- Spearman, Charles. ” general intelligence,” objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.
- Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Srivastava, Rupesh Kumar, Greff, Klaus, and Schmidhuber, Jürgen. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.
- Sun, Chen, Shetty, Sanketh, Sukthankar, Rahul, and Nevatia, Ram. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pp. 371–380. ACM, 2015.
- Tang, Yichuan. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.
- Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5 - rmsprop., 2012.
- Vinyals, Oriol, Kaiser, Lukasz, Koo, Terry, Petrov, Slav, Sutskever, Ilya, and Hinton, Geoffrey. Grammar as a foreign language. *arXiv preprint arXiv:1412.7449*, 2014.
- Zhang, Saizheng, Wu, Yuhuai, Che, Tong, Lin, Zhouhan, Memisevic, Roland, Salakhutdinov, Ruslan, and Bengio, Yoshua. Architectural complexity measures of recurrent neural networks. *arXiv preprint arXiv:1602.08210*, 2016.